ON MEASURING OF SIMILARITY BETWEEN TREE NODES

Gleb B. Sologub
Moscow Aviation Institute (State Technical University)
e-mail: glebsologub@ya.ru

Abstract

In this paper, a survey of similarity measures between vertices of a graph is presented. Distance-based and structural equivalence measures are described. It is demonstrated that most of them degenerate if applied directly to the tree nodes. Adjusted path-based similarity measure is proposed as well as a new method for representing tree nodes as binary vectors that is based on using of an ancestor matrix. It is shown that application of ordinary similarity measures to this representation gives desired non-trivial results.

Keywords: *Similarity measure, distance on tree nodes, structural equivalence, ancestor matrix.*

1. INTRODUCTION

The concept of similarity is commonly used in relation with clustering and collaborative filtering methods in many fields, including information retrieval, data mining, network analysis, pattern recognition and machine learning. Basic task for these methods is to calculate similarity between data entries and find most similar to one another.

Tree structures are used to represent various types of hierarchical data. Examples include different ontologies, catalogs, genealogies, XML documents, language corpuses, etc.

In our work on intelligent tutoring and testing systems we need to evaluate similarity between the questions of a test in order to predict answer scores. We use tree data structures for domain modeling. Nodes of a tree represent themes or subjects; leaves represent questions. So, the main goal of our study is to develop effective and accurate measure of similarity between tree leaves.

In this paper, following the work [1], we discuss only abstract graph theoretic methods to compute the similarity on tree nodes without any regard to the problem domain.

Perfect studies on different approaches to the measuring of similarity as semantic distance that do relate to the problem domain, i.e. information retrieval, could be found in the works [2], [3], and [4].

2. PRELIMINARIES

A tree is a connected undirected simple graph with no cycles. Any two nodes of a tree are connected by a unique simple path, which is the shortest path between them. We consider a rooted tree, which has a root node and leaves.

We denote the number of tree nodes by n, nodes (vertices) by $v_1, v_2, \ldots$; particularly, root node by t, leaves by $q_1, q_2, \ldots$; parent nodes by $t_1, t_2, \ldots$; the lowest common ancestor of vertices v_i and v_j by lca_{ij}, length of the shortest path between vertices v_i and v_j by $l(v_i,v_j)$, number of common neighbors of vertices v_i and v_j by n_{ij}.

Also we use the following notation: **A** for adjacency matrix, a_{ij} for its elements, A_i for its rows, A_j for its columns, **I** for identity matrix, $\tilde{\mathbf{A}}$ for ancestor matrix with elements $\tilde{a}_{i,j}$=1 iff the jth vertex is an ancestor of the ith vertex, k_i for degree of ith vertex, **D** for diagonal degree matrix with elements $d_{ii}=k_i$, **L** for Laplacian matrix, which is **D**–**A**.

Note that $a_{ij}=a_{ji}=\{0 \text{ or } 1\}$ and $a_{ij}^2=a_{ij}$ for all $i, j; n_{ij} = \sum_k a_{ik}a_{kj}$, $k_i = \sum_k a_{ik}$.

3. DISTANCE ON VERTICES

Similarity is somewhat opposite to the concept of distance between information elements. One can use distances or metrics to construct similarity measure for any kinds of elements. For example, if $d(x,y)$ is a distance between x and y, then their similarity could be measured as follows [5]:

$$s(x,y) = \frac{1}{1+d(x,y)}. \qquad (1)$$

In general, many types of monotonically decreasing functions could be used for this purpose.

3.1. Path metric

The obvious measure for distance on tree nodes could be a path metric [6], i.e. length of the shortest path between them:

$$l(v_i,v_j)=l(v_i,lca_{ij})+l(v_j,lca_{ij}) \qquad (2)$$

The similarity measure that is based on path metric then could be expressed as

$$s_l(v_i,v_j) = \frac{1}{1+l(v_i,v_j)} = \frac{1}{1+l(v_i,lca_{ij})+l(v_j,lca_{ij})}. \qquad (3)$$

But it is not so useful for hierarchical data structure, because it makes no difference between similarities of node pairs located at different depths.

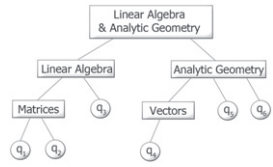

Figure 1. Example of a simple curriculum

Consider a simple curriculum (Figure 1). It is obvious that the similarity between questions q_1 and q_2 should be greater than the similarity between questions q_5 and q_6, because they belong to the more specific theme "Matrices". However, the distances between them are equal.

Later, we shall improve path-based similarity measure by removing this effect.

3.2. Resistance distance

The resistance distance $\Omega_{i,j}$ between vertices v_i and v_j of a simple connected graph G could be used to compute similarity [7] and is defined as

$$\Omega_{i,j}=\Gamma_{i,i}+\Gamma_{j,j}-\Gamma_{i,j}-\Gamma_{j,i}. \quad (4)$$

where Γ is the Moore-Penrose inverse of the Laplacian matrix **L** of G.

However, it is shown [8] that in the case of a tree:

$$\Omega_{i,j}=\det \mathbf{L}[i;j]=l(v_i,v_j), \quad (5)$$

where $\mathbf{L}[i;j]$ is a submatrix of **L** that is obtained by deleting the ith and the jth rows and columns from **L**.

Sad, but the resistance distance in a tree is just a path metric again.

3.3. Adjusted path-based similarity

Now return to the path metric. A simplest way to account for the granularity of the domain, to which belong concerned vertices, is to adjust formula (2) as

$$l_a(v_i,v_j)=\frac{l(v_i,lca_{ij})+l(v_j,lca_{ij})}{1+l(lca_{ij},t)}. \quad (6)$$

Obviously, l_a is not a metric. This could be simply illustrated by a contrary instance.

For example, in Figure 2, $l_a(v_i,v_k)=6$, $l_a(v_i,v_j)=4/3$, $l_a(v_j,v_k)=4$, so $l_a(v_i,v_k)>l_a(v_i,v_j)+l_a(v_j,v_k)$.

Nevertheless we could use la as a dissimilarity measure, since it is larger for vertices that are more distant to each other.

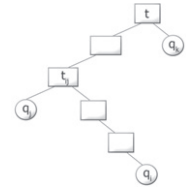

Figure 2. Example of a tree, where $l_a(v_i,v_k)>l_a(v_i,v_j)+l_a(v_j,v_k)$

So the adjusted path-based similarity measure could be written as

$$s_a(v_i,v_j)=\frac{1}{1+l_a(v_i,v_j)}=\frac{1+l(lca_{ij},t)}{1+l(lca_{ij},t)+l(v_i,lca_{ij})+l(v_j,lca_{ij})} \quad (7)$$

4. STRUCTURAL EQUIVALENCE

Two vertices of a graph are called structurally equivalent if they share the same neighbors. Thus, the similarity of vertices could be expressed by generalization of the number of common neighbors.

As the simplest and most obvious measure for the structural equivalence, the number of common neighbors is used itself [1].

But in the case of a tree it turns to be a binary variable that is equal to 1 if two vertices have the same parent, and is equal to 0 otherwise. So it is almost useless value.

4.1. Cosine similarity

One of the most popular similarity measures is a cosine similarity. It is defined by the following simple formula [9]:

$$\sigma_{ij} = \cos\theta = \frac{(x,y)}{\|x\|\,\|y\|}, \tag{8}$$

where x and y are two vectors, $\|x\|$ and $\|y\|$ are the norms of x and y, (x,y) is their dot product and θ is the angle between them.

It is often proposed to represent vertices of a graph as corresponding rows (or columns) of the adjacency matrix, so we could obtain that [1]:

$$\sigma_{ij} = \frac{\sum_k a_{ik}a_{kj}}{\sqrt{\sum_k a_{ik}^2}\sqrt{\sum_k a_{jk}^2}} = \frac{n_{ij}}{\sqrt{k_i k_j}}. \tag{9}$$

This value is almost useless again in the case of tree nodes. Especially, this is true for tree leaves, because they always have degree 1.

4.2. Euclidean distance

Given two vectors x and y we could compute the Euclidean distance between them:

$$\rho_E = \|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}. \tag{10}$$

For the distance on graph nodes it could be written as

$$\rho_E(A_i,A_j) = \sqrt{\sum_k (a_{ik} - a_{jk})^2} = \\ = \sqrt{\|a_i\|^2 + \|a_j\|^2 - 2(a_i,a_j)} = \sqrt{k_i + k_j - 2n_{ij}}, \tag{11}$$

This formula gives another degenerated measure on nodes and, especially, leaves of a tree.

4.3. Tanimoto similarity measure

The next similarity measure that deals with vectors is the Tanimoto coefficient [9]:

$$S_T = \frac{(x,y)}{\|x\|^2 + \|y\|^2 - (x,y)}, \tag{12}$$

or using the previous representation of graph vertices as rows of adjacency matrix:

$$S_T(A_i,A_j) = \frac{n_{ij}}{k_i + k_j - n_{ij}}. \tag{13}$$

It is a different mix of degrees and common neighbor counts that gives trivial results on tree nodes and leaves.

Consider two sets M and N. Jaccard index [6] is defined on these two sets as

$$J(M,N) = \frac{|M \cap N|}{|M \cup N|} = \frac{|M \cap N|}{|M| + |N| - |M \cap N|}. \quad (14)$$

Jaccard index measures the similarity between two given sets as the size of their intersection divided by the size of their union.

Let us arrange all members of $|M \cap N|$ in an ordered list L with elements l_i. Consider binary vectors x and y with respective components:

$$x_i = \begin{cases} 1, & if\ l_i \in M \\ 0, & otherwise \end{cases}, and\ y_i = \begin{cases} 1, & if\ l_i \in N \\ 0, & otherwise \end{cases}. \quad (15)$$

Tanimoto coefficient of these vectors is equivalent to the Jaccard index of the given sets [9]:

$$S_T(x,y)=J(M,N). \quad (16)$$

4.4. Pearson coefficient

One could use the standard Pearson correlation coefficient as the measure of similarity between two given vertices as:

$$r_{ij} = \frac{\operatorname{cov}(A_i, A_j)}{\sigma_i \sigma_j} = \frac{n_{ij} - \frac{k_i k_j}{n}}{\sqrt{k_i - \frac{k_i^2}{n}}\sqrt{k_j - \frac{k_j^2}{n}}} = \frac{n_{ij} n - k_i k_j}{\sqrt{k_i n - k_i^2}\sqrt{k_j n - k_j^2}}. \quad (17)$$

And again, for leaves of a tree, we obtain degenerated formula

$$r_{ij} = \frac{n_{ij} n - 1}{n - 1} = \begin{cases} 1, & if\ v_i\ and\ v_j\ have\ same\ parent \\ -\frac{1}{n-1}, & otherwise \end{cases} \quad (18)$$

4.5. Different representation of tree vertices

Other kinds of measures could be applied to binary vectors. Examples include various weighted metrics, set and string distances, and even logical comparison [9]. But if we remain to use them directly on rows of adjacency matrix of a tree, then the results will be trivial again.

We propose another way to represent tree nodes that is based on using of an ancestor matrix instead of the adjacency matrix. The ancestor matrix $\tilde{\mathbf{A}}$ of a graph is defined as a square matrix where an element $\tilde{a}\neg_{i,j}$ is set to 1 if the jth vertex is an ancestor of the ith vertex, and 0 otherwise. The ancestor matrix is less sparse than the adjacency matrix of a tree, so it gives us more effect.

It should be noted that different vertices v_i and v_j of a graph can have equal corresponding rows A_i and A_j of its adjacency matrix. Particularly, this applies to any pair of leaves, such that they are children of the same parent node in a tree. Thus any of the similarity measures, described in this chapter, would give the highest value on this pair of leaves. Such behavior is undesirable, because we assume that only identical elements should have the highest value of similarity [3]. This could be observed in the case of using of $\tilde{A}_i$ and $\tilde{A}_j$ too.

To get rid of this effect we propose to use rows of **C=I+Ã** matrix as binary vectors for measuring of distances and similarity between nodes of a tree.

Two following results of this approach reveal the relationship between the graph distance on tree nodes and the metric on rows of the extended ancestor matrix **C** of this tree.

Theorem 1. Let T be a rooted tree with ancestor matrix **Ã**. Then

$$s_a(v_i,v_j)=S_T(C_i,C_j) \tag{19}$$

for any two vertices v_i, v_j of T and corresponding rows C_i, C_j of **C=I+Ã**.

Proof: Consider the sets $P_i=\{v_i,t_{i_1},t_{i_2},...,lca_{ij},t_{k_1},t_{k_2},...,t\}$ and $P_j=\{v_j,t_{j_1},t_{j_2},...,lca_{ij},t_{k_1},t_{k_2},...,t\}$, where t is the root of T, lca_{ij} is the lowest common ancestor of v_i and v_j in T, t_{k_p} are their other common ancestors; t_{i_l}, t_{j_m} are the other ancestors of given vertices v_i and v_j, respectively. In this notation, using equation (16) and definition of the Jaccard index (14) we directly obtain that

$$S_T(C_i,C_j)=J(P_i,P_j)=\frac{|P_i\cap P_j|}{|P_i|+|P_j|-|P_i\cap P_j|}. \tag{20}$$

We recall that the length of the shortest path between two vertices is one less than the number of vertices in this path. Then we notice that

$$(C_i,C_j)=|P_i\cap P_j|=|\{lca_{ij},t_{k_1},t_{k_2},...,t\}|=1+l(lca_{ij},t), \tag{21}$$

$$\|C_i\|^2=|P_i|=1+l(v_i,t)=1+l(v_i,lca_{ij})+l(lca_{ij},t), \tag{22}$$

$$\|C_j\|^2=|P_j|=1+l(v_j,t)=1+l(v_j,lca_{ij})+l(lca_{ij},t). \tag{23}$$

Finally, we can write

$$S_T(C_i,C_j)=$$

$$=\frac{1+l(lca_{ij},t)}{1+l(v_i,lca_{ij})+l(lca_{ij},t)+1+l(v_j,lca_{ij})+l(lca_{ij},t)-(1+l(lca_{ij},t))}. \tag{24}$$

It turns exactly to the sa(vi,vj) by some trivial algebra. ▪

Corollary 1. We can define a proper metric on vertices of T as

$$\tilde{l}_a(v_i, v_j) = \frac{l(v_i, v_j)}{1 + l(lca_{ij}, t) + l(v_i, v_j)}. \quad (25)$$

Proof: Formula (25) is derived immediately by defining

$$\tilde{l}_a(v_i, v_j) = 1 - s_a(v_i, v_j)$$

from

$$1 - s_a(v_i, v_j) = 1 - \frac{1 + l(lca_{ij}, t)}{1 + l(lca_{ij}, t) + l(v_i, lca_{ij}) + l(v_j, lca_{ij})} =$$
$$= \frac{l(v_i, lca_{ij}) + l(v_j, lca_{ij})}{1 + l(lca_{ij}, t) + l(v_i, lca_{ij}) + l(v_j, lca_{ij})}.$$

Theorem 1 shows that $1-s_a(v_i,v_j)$ is equal to the Tanimoto distance $1-S_T(C_i,C_j)$, and the Tanimoto distance is known to be a proper metric [9]. ▪

Theorem 2. Within notation of Theorem 1,

$$\rho_E(C_i, C_j) = \sqrt{l(v_i, v_j)}. \quad (26)$$

Proof: Using the same approach and formulas (11), (21), (22) and (23) we obtain

$$\rho_E(C_i, C_j) = \quad (27)$$
$$= \sqrt{1 + l(v_i, lca_{ij}) + l(lca_{ij}, t) + 1 + l(v_j, lca_{ij}) + l(lca_{ij}, t) - 2(1 + l(lca_{ij}, t))},$$

which is equal to

$$\sqrt{l(v_i, lca_{ij}) + l(v_j, lca_{ij})} = \sqrt{l(v_i, v_j)}. \blacksquare$$

For other kinds of metrics and similarity measures, defined on rows of the extended ancestor matrix, path-based expressions could be obtained by using the same technique.

5. DISCUSSION

Similarity measures on tree nodes were discussed primarily in relation with semantic similarity and its applications [2, 3, and 4]. Edge-counting methods were well developed in this area. The closest form to our adjusted path-based similarity measure is that was proposed in [10]. It will be interesting to adopt the technique shared by previous researchers [2, 3, and 4] to compare these measures in terms of correlation with human judgment.

Some of tree comparison methods, e.g. consensus methods, are based on computing of similarity between tree nodes too. Moreover, in the particular work [11] a set similarity measure in the form of Jaccard index is used. The difference is that they define it on leaf sets under nodes of leaf-labeled trees, whereas we consider extended ancestor sets for nodes of any rooted tree.

While the theoretical relationship between resistance distance in graph and Euclidean distance in some vector space is well known [8], we believe that particular result, obtained in Theorem 2, was not observed earlier.

Theorem 1 and Theorem 2 give us a way to compute distances on nodes of a tree using standard vector operations.

On the other hand they provide us with simple path-based methods to measure similarity in tree structured data.

We propose to use these measures in many other related areas, for example, content-based image retrieval [12] or case-based reasoning student diagnosis [13].

6. CONCLUSION

This work provides a survey of similarity measures on nodes of a tree. Distance-based and structural equivalence measures are discussed. A new method for representing tree nodes and its use for measuring similarity is described. Theorems 1 and 2 give interesting results about relationship between paths on tree nodes and metrics on rows of extended ancestor matrix of this tree. Future work would be related with further studying of different similarity measures and their comparative analysis.

REFERENCES

1. **Newman, M.E.J.** 2010. Networks: An Introduction. Oxford University Press.
2. **Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E.G.M., Milios, E.** 2006. Information Retrieval by Semantic Similarity. Intern. Journal on Semantic Web and Information Systems (IJSWIS), 3(3), July/Sept. 2006, 55–73.
3. **Lin, D.** 1998. An Information-Theoretic Definition of Similarity. In Proc. of the 15th Int. Conference on Machine Learning, 296-304.
4. **Jiang, J.J., Conrath, D.W.** 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In Proc. of Int. Conference Research on Computational Linguistics (ROCLING X), 1997, Taiwan.
5. **Segaran, T.** 2007. Programming Collective Intelligence. O'Reilly Media.
6. **Deza, M.M., Deza, E.** 2009. Encyclopedia of Distances. Springer.
7. **Kunegis, J., Schmidt, S., Albayrak, S., Bauckhage, C., Mehlitz M.** 2008. Modeling Collaborative Similarity with the Signed Resistance Distance Kernel. In Proc.European Conf. on Artificial Intelligence, 261-265.
8. **Klein, D.J., Randić, M.** 1993. Resistance distance. Journal of Mathematical Chemistry, V. 12, N. 1, 81-95.
9. **Kohonen, T.** 2001. Self-Organizing Maps. Springer.
10. **Wu, Z., Palmer, M.** 1994. Verb semantics and lexical selection. In Proc. of the 32nd Annual Meeting of the Associations for Computational Linguistics, Las Cruces, New Mexico, 133–138.
11. **Zhang, L.** 2003. On matching nodes between trees. Tech. Rep. N. 2003-2067. HP Labs.

12. **Manouvrier, M., Rukoz, M., Jomier, G.** 2005. A generalized metric distance between hierarchically partitioned images. In Proc. of the 6th Int. Workshop MDM/KDD'05, August 21, 2005, 33-41.
13. **Tsaganou, G., Grigoriadou, M., Cavoura, T.** 2002. Case-based reasoning diagnosis of students' cognitive profiles on historical text comprehension. In Proc. IEEE Int. Conf. on Advanced Learning Technologies (ICALT 2002).